Assignment 4

Ariful Islam

June 2025

1 Introduction

We performed an adversarial attack experiment using a horned viper image on a pre-trained MobileNetV2 model trained on the ImageNet dataset. For the attack, we used the Fast Gradient Sign Method (FGSM) with three different epsilon values: 0.01, 0.10, and 0.15. At first in 0.01 the model increase the confidence score with a little noise, but when the epsilon increased, the model's prediction confidence decreased and the predicted class changed. For example, at epsilon 0.10 the model misclassified the image as a sea cucumber, and at 0.15 it predicted peacock, which proves that FGSM can successfully done the model even with small noise.

Then, we tested Gaussian noise by adding random values to the input image with the same epsilon. Visually the image looked noisy, but the model still predicted a horned viper with high confidence (85.87 percent). This proves that Gaussian noise is not effective as an adversarial attack, because it is not targeted and doesn't follow the model's gradients like FGSM does.

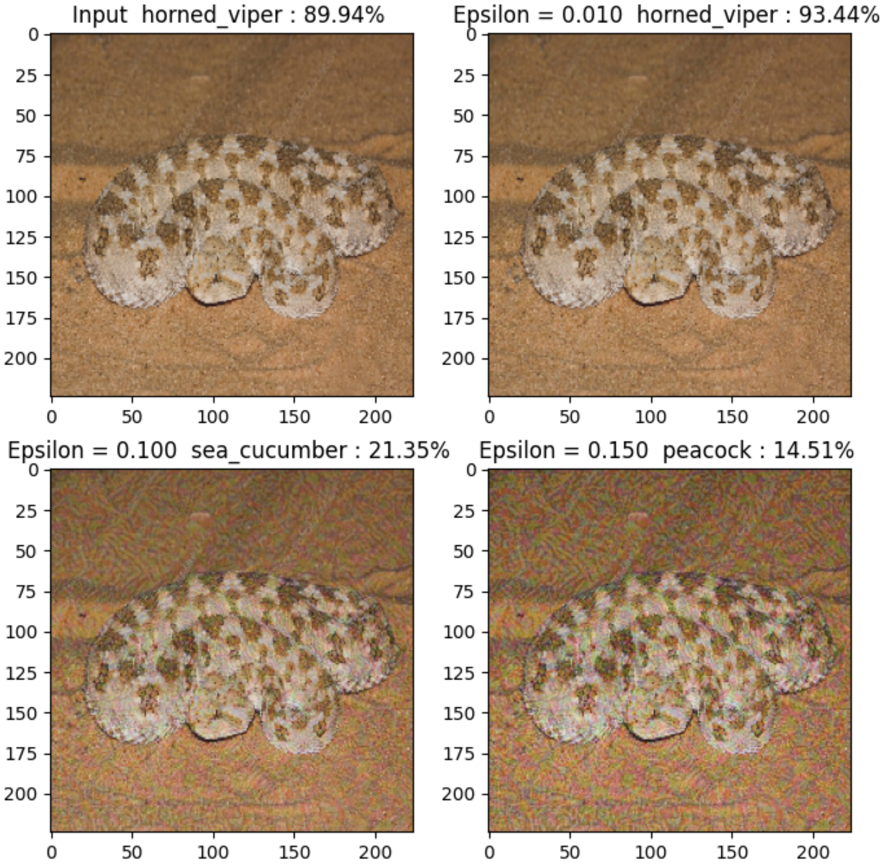
Input horned_viper : 89.94%
Epsilon = 0.010 horned_viper : 93.44%
Epsilon = 0.100 sea_cucumber : 21.35%
Epsilon = 0.150 peacock : 14.51%
0
25
50
75
100
125
150
175
200
0
50
100
150
200

Figure 1: The result of FGSM.

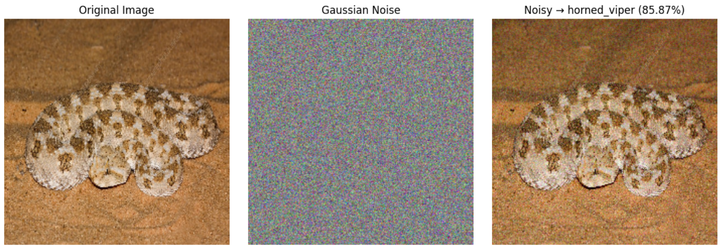

Figure 2: The result of adding Gaussian noise to see whether the FGSM works or not.